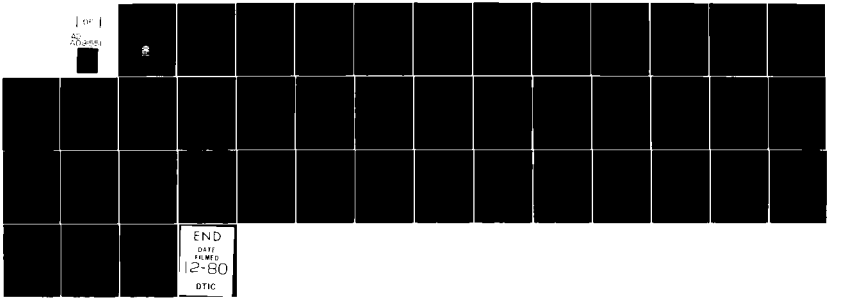
AD-A091 551 SOUTHERN METHODIST UNIV DALLAS TEX DEPT OF STATISTICS F/G 12/1
BIASED REGRESSION: A TEN YEAR PERSPECTIVE,(U)
1980 R F GUNST
UNCLASSIFIED F49620-79-C-0106
NL



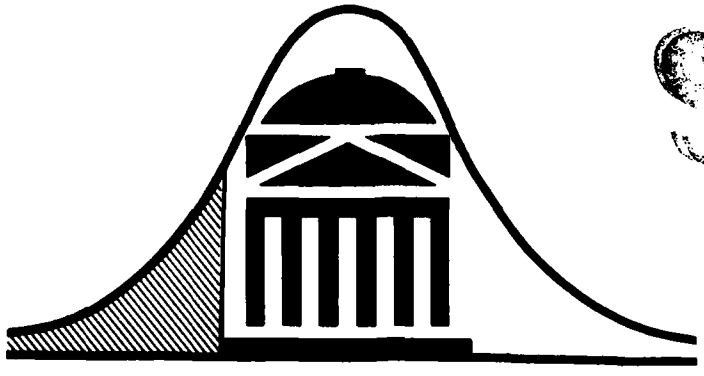
END
DATE FILMED
12-80
DTIC

SOUTHERN METHODIST UNIVERSITY

BIASED REGRESSION:
A TEN YEAR PERSPECTIVE

Richard F. Gunst

Department of Statistics
Southern Methodist University
Dallas, Texas 75275



DEPARTMENT OF STATISTICS

DALLAS, TEXAS 75275

BIASED REGRESSION: A TEN YEAR PERSPECTIVE

Richard F. Gunst
Department of Statistics
Southern Methodist University
Dallas, Texas 75275

ABSTRACT

Biased estimation in regression has experienced a tremendous growth in popularity since Hoerl and Kennard's formalization of ridge regression. Along with the interest in biased regression, many research efforts over the last ten years have extended both the theory and application of these methodologies. So too, criticisms have arisen which focus on the incompleteness of the theoretical results and on exaggerated claims about the merits of biased estimators. Rather than attempting to arbitrate these opposing views, this article discusses biased estimation with special emphasis on its ultimate justification: application to real problems. Advantages and disadvantages of three biased estimators (principal component, latent root, and ridge regression estimators) are discussed and illustrated through a comprehensive analysis of a data set on automobile emissions. From the discussion and analysis it is hoped that a more balanced perspective on the application of biased estimation will be fostered.

1. INTRODUCTION

One of the most important advances in statistical methodology over the last ten years has been the introduction and refinement of biased regression techniques. Biased regression is not altogether new; some procedures (e.g. principal component regression) were popular long before 1970. Nevertheless, at the outset of the last decade least squares was the (almost) universally accepted estimation scheme. The major impact of the newer biased regression methodologies and the resurgence of some older ones has been their widespread acceptance and utilization as alternatives to least squares by a very large community of regression analysts in a variety of fields of application.

So widespread is the application of biased regression that the euphoria surrounding its successful implementation has generated exaggerated claims concerning the benefits that can be realized with biased estimators. Major criticisms of biased regression are beginning to surface, criticisms which are necessary for a proper perspective of the role of biased estimation in a regression analysis. Some of the criticism, however, is based as much on the exaggerated claims of users of the methodologies as on the true faults of the techniques. Indeed, some of the current criticism is supported neither by theoretical arguments nor by the views or intentions of the original authors. Thus both supporters and critics of biased regression are guilty of arguing from unsubstantiated viewpoints.

This article will review three of the more popular biased estimators which have been introduced or refined over the last ten years: principal component regression, latent root regression, and ridge regression. The scope of the paper is intentionally limited to these estimators and to a

discussion of their strengths and weaknesses. Shrunken estimation, minimax estimators, Bayesian regression and robust regression are additional estimation schemes which achieved major advances over the last decade but the inclusion of which would render this discussion overly broad and unwieldy. Throughout this article, then, the term "biased regression" will refer generically to the three specific biased estimators mentioned above.

2. BIASED REGRESSION AND LEAST SQUARES

Biased regression estimators were not proposed as replacements for least squares estimators. Draper and Van Nostrand (1979, p. 464) are correct in insisting that "The extended inference that ridge regression is 'always' better than least squares is, typically, completely unjustified." The same type of criticism can be levied at all regression estimators, including least squares: no regression estimator has been shown to be universally superior according to all reasonable criteria of evaluation.

Another point which must be understood when assessing the merits of biased estimators is that they were not developed solely in order to produce regression estimators that have better theoretical properties than least squares estimators. To be sure, biased estimators which can be shown to have decidedly inferior theoretical properties should be discarded, but the "reasonable criteria" referred to above must include the judgement and insight of the investigator. Hoerl (1962) proposed the ridge estimator as a means of altering least squares estimates which were found to be unacceptable from a chemical engineering standpoint almost a decade before Hoerl and Kennard (1970a) buttressed the value of ridge regression with theoretical arguments.

This facet of the literature on biased regression is often overlooked when attempts are made to assess biased methodologies. Many of the advances in biased estimation were stimulated because of the inability of least squares to produce acceptable estimates on specific real data sets. Hoerl and Kennard (1970a) begin their derivation of the ridge estimator by citing the occurrence of such problems: "The least squares estimates often do not make sense when put into the context of the physics, chemistry, and engineering of the process which is generating the data." While it is essential that theoretical properties of biased estimators be established to insure they are not just ad-hoc replacements to least squares, a major motivation for the advocacy of biased regression is the realization that properties of the data sometimes preclude satisfactory estimation with least squares.

One property of the data which often results in unacceptable least squares estimates is that of multicollinearity. The detrimental impact of multicollinearities on least squares estimators are examined in Gunst and Mason (1977a), Hoerl and Kennard (1970a), and Silvey (1969) and will only be outlined here. Write the regression model as

$$\underline{Y} = \beta_0 \underline{1} + X\underline{\beta} + \underline{\varepsilon} \tag{2.1}$$

where $\underline{Y}$ is an (n×1) vector of observable response variables, $\underline{1}$ is a vector of ones, X is an (n×p) full-column-rank matrix of standardized (X'X is in correlation form) nonstochastic predictor variables, $\underline{\varepsilon}$ is a vector of unobservable error terms with $\varepsilon_i \sim NID(0, \sigma^2)$, and β_0 and $\underline{\beta}$ are the unknown constant term and vector of regression coefficients. The least squares estimator of $\underline{\beta}$ is

$$\hat{\underline{\beta}}_{LS} = (X'X)^{-1}X'\underline{Y}. \tag{2.2}$$

Let the latent roots and latent vectors of X'X be denoted by $0<\ell_1\leq\ell_2\leq \ldots \leq\ell_p$ and $\underline{V}_1, \underline{V}_2, \ldots, \underline{V}_p$, respectively. Then

$$\text{(i) } E[\hat{\underline{\beta}}_{LS}'\hat{\underline{\beta}}_{LS}] = \underline{\beta}'\underline{\beta} + \sigma^2 \sum_{j=1}^{p} \ell_j^{-1}$$

(ii) $\text{Var}[\underline{V}_j'\hat{\underline{\beta}}_{LS}] = \sigma^2/\ell_j$

(iii) $\hat{\underline{\beta}}_{LS} = \sum_{j=1}^{p} \ell_j^{-1} c_j \underline{V}_j$, where $c_j = \underline{V}_j'X'\underline{Y}$.

Properties (i) to (iii) illustrate three commonly-noticed difficulties with least squares estimators. If multicollinearities exist in X, one or more of the latent roots of X'X will be close to zero. Then (i) asserts that the length of the least squares estimator will tend to be much greater than that of $\underline{\beta}$ unless σ^2 is very small, (ii) reveals that some linear combinations of $\underline{\beta}$ will be estimated with much worse precision than others, and (iii) indicates that the magnitudes (through large ℓ_j^{-1}) and signs (through the signs of the elements of $\underline{V}_j$) can be greatly affected by multicollinearities in the data (see the above references for further discussion on these properties). Of course, a particular data set might not exhibit detrimental properties similar to these and least squares can yield acceptable results. The message articulated above by Hoerl and Kennard (and others) is that, in their experience, these deleterious effects are often observed when least squares is utilized in conjuntion with multicollinear predictor variables.

Gunst and Mason (1977b), Hoerl and Kennard (1970b), and Marquardt and Snee (1975) analyze data sets which exhibit one or more of the preceeding difficulties and for which the resulting least squares estimates appear to be inconsistent with intuitive or known physical characteristics of the respective studies. For completeness and later reference, a further illustration is now provided utilizing data contained in Hare and Bradow (1977). In this study the investigators sought to obtain correction factors in order to adjust diesel truck emissions for ambient humidity, atmospheric pressure, and temperature levels. By using regression models, emissions readings could

be adjusted to "standard" humidity, pressure, and temperature levels. The regression analysis examined here consists of 44 experiments conducted on a single diesel truck. In each experiment temperature was controlled but humidity and pressure conditions were not. Table 1 displays three least squares fits of these variables to nitrous oxides (NO_x) emitted by the engine.

[Insert Table 1]

Prior to the analysis, the investigators were not able to specify the precise form of the regression model but believed that increases in humidity would be associated with substantial decreases in NO_x and that changes in temperature would be associated with relatively minor changes in NO_x (Hare and Bradow, p. 2571). In Table 1, three fits to the NO_x emissions are shown. The first fit includes only the linear terms for humidity (H), pressure (P), and temperature (T), the second one adds interaction terms to the linear ones, and the third one also includes the quadratic terms. The coefficients of determination (R^2) indicate that the latter two fits offer substantial improvement to the linear fit and that the full quadratic fit appears to be a reasonable improvement over the interaction terms. Disturbing, however, are the changes in signs and magnitudes of the coefficient estimates as terms are added to the linear fit.

In the linear fit, the coefficient on humidity is relatively large and negative while the one on temperature is much smaller - both of which were anticipated by the investigators. As the interaction and quadratic terms are added the signs and magnitudes of the estimates fluctuate wildly with humidity and pressure increasing dramatically in magnitude and that for temperature first increasing dramatically (in magnitude) and then dropping just as rapidly. Note especially that in the quadratic fit only the H×P coefficient is negative (and no other coefficient involving humidity) and that H, P, H×P, and P^2

Table 1. Initial Least Squares Fits to NO_x Emissions Data

Predictor Variable	Coefficient Estimates: Linear Fit	Linear and Interactions	Quadratic Fit	$\ell_1 = .517\times10^{-6}$ $\underline{V}_1$	$\ell_2 = .196\times10^{-5}$ $\underline{V}_2$
Humidity (H)	-.240	6.841	44.738	-.529	.061
Pressure (P)	.394	-2.090	69.736	-.448	-.313
Temperature (T)	.055	-29.215	.734	.125	-.642
P×T		28.090	- 1.710	-.129	.631
H×P		-12.010	-46.019	.531	-.118
H×T		5.309	.797	-.017	.077
P^2			-69.005	.453	.261
T^2			.971	.009	-.010
H^2			.458	.009	-.012
R^2	.695	.795	.835		

have coefficients that are over two orders of magnitude larger than any of those in the linear fit.

As one might surmise, the least squares estimates fluctuate wildly due to correlations among the humidity and pressure readings. This fact was unanticipated prior to the experiment but explained afterward as ascribable to local environmental conditions: "decreases in atmospheric pressure are frequently followed by southerly winds carrying humid air from the Gulf of Mexico." Thus, while these conditions would not routinely be expected to exist in other locals, during the experiment the correlation between H and P was -0.816. Because of this correlation and because atmospheric pressure did not vary greatly over the duration of the experiment, the correlation between H and H×P was greater than 0.999, that between P and P^2 was also greater than 0.999, and that between T and P×T was 0.996.

These three large pairwise correlations are also identifiable in the latent vectors which correspond to the two smallest latent roots of X'X for the full quadratic fit. Displayed in Table 1, the two smallest latent roots are very close to zero. The four largest magnitudes in $\underset{\sim}{V}_1$ correspond to H, P, H×P, and P^2 and the two large ones in $\underset{\sim}{V}_2$ correspond to T and P×T. Note too that in $\underset{\sim}{V}_1$ the magnitudes of H and H×P are about equal and opposite in sign, those of P and P^2 are about equal and opposite in sign, and those of T and P×T in $\underset{\sim}{V}_2$ are about equal and opposite in sign. This occurrence is characteristic of strong (positive) pairwise multicollinearities. Their effect on least squares estimates is likewise characteristic.

Observe that the four least squares estimates with the largest magnitudes in the quadratic fit correspond to the four large elements in $\underset{\sim}{V}_1$. Note too that the coefficient estimates for H and H×P are about equal and opposite

in sign, as are those of P and P^2 and to a slightly lesser extent those of T and P×T. That the signs on the estimates are not identical to those in $\underline{V}_1$ and $\underline{V}_2$ is unimportant since the latent vectors are only unique up to a multiple of -1. It is apparent that the magnitudes and signs of the estimates are being determined by the multicollinearities among the predictor variables.

Two suggestions are often made concerning the possible rectification of the problems just discovered. One is that more data should be collected, data which do not exhibit these same multicollinearities. In the present instance the suggestion might be impossible to accomplish. Even if additional time and resources could be provided for continuation of the experiment, the local climatic conditions will inevitably produce the same correlations between humidity and pressure as well as the narrow range of pressure readings. Perhaps all the equipment and personnel could be moved to a new location but this would be at a major cost to the contractor. Regardless of whether new experimentation is possible, one cannot merely suggest that the only viable solution is to collect more data without regard to the feasibility or ramifications of such a suggestion.

A second position that is often argued is that better use should be made of a priori information in the estimator itself. Whenever the information is of sufficient refinement that it can be so incorporated, we wholeheartedly agree. A priori information is not always so refined, however, and is often very crude. In this analysis one could employ estimators which restrict the coefficient on humidity to be negative, consistent with the investigators' a priori assessment of the effect of humidity on NO_x. It is more difficult and subjective to mathematically specify the relatively

lower importance of temperature. In addition, it is not always clear how this information should generalize to models with interactions and quadratics. Finally, the correlation between humidity and pressure, while explained afterwards, was not anticipated prior to the experiment. Often such a correlation is not explainable following an experiment but nevertheless exists and must be accounted for in the regression estimator. A variety of estimators can do so.

To summarize this section, biased estimators are not intended to replace least squares. Least squares estimators have valuable theoretical and empirical properties as well as a wealth of ancillary techniques associated with them for variable selection, outlier detection, model assessment, etc. Least squares will continue to be the single most important and popular regression methodology. With multicollinear data, however, least squares estimates are frequently at odds with known or suspected properties of the model. The cause of the difficulties can usually be traced to the multicollinearities themselves. Several remedial strategies are then available to the regression analyst, one of which is biased estimation.

3. BIASED REGRESSION ESTIMATORS

Since the latent vectors of X'X span p-space, the parameter vector $\underline{\beta}$ can always be expressed as

$$\underline{\beta} = \sum_{j=1}^{p} \alpha_j \underline{v}_j \tag{3.1}$$

for appropriately chosen constants $\alpha_1, \alpha_2, \ldots, \alpha_p$. Ultimately, then, estimation of $\underline{\beta}$ consists of selecting suitable values for the α_j in eqn. (3.1). In the last section it was noted that

$$\hat{\underline{\beta}}_{LS} = (X'X)^{-1} X'\underline{Y} = \sum_{j=1}^{p} \ell_j^{-1} c_j \underline{v}_j, \tag{3.2}$$

where $c_j = \underline{V}_j'\underline{X}'\underline{Y}$. Because of the large values of ℓ_j^{-1} for latent vectors identifying multicollinearities, least squares estimators allow the first few terms in eqn. (3.2) to dominate the estimation of the regression coefficients, at least for those coefficients corresponding to multicollinear predictor variables. Note that the large values of ℓ_j^{-1} are mandated by the strength of the multicollinearities and are in no way influenced by the relative magnitudes of the α_j in eqn. (3.1).

All three biased estimators discussed below dampen the least squares weights on the $\underline{V}_j$ in eqn. (3.2). In doing so, some of the deleterious effects of multicollinearities on the least squares estimators can be alleviated. In fact, all three biased estimators can be derived as least squares estimators subject to restrictions on the effects of the multicollinearities (see Hocking (1976)). In the following subsections, each of the biased estimators is derived and its benefits outlined.

3.1 Principal Component Regression

From eqn. (3.2), observe that $\underline{V}_j'\hat{\underline{\beta}}_{LS} = \ell_j^{-1}c_j$. One technique for deriving the principal component estimator, $\hat{\underline{\beta}}_{PC}$, is to minimize the residual sum of squares $(\underline{Y} - \hat{\underline{Y}})'(\underline{Y} - \hat{\underline{Y}})$ subject to the restrictions $\underline{V}_j'\hat{\underline{\beta}} = 0$ for some subset of the latent vectors of X'X. Such an estimator completely eliminates the effect of the latent vectors in the subset on the estimation of $\underline{\beta}$. Thus, $\hat{\underline{\beta}}_{PC}$ is chosen to minimize

$$\phi = (\underline{Y} - \hat{\beta}_0\underline{1} - \underline{X}\hat{\underline{\beta}})'(\underline{Y} - \hat{\beta}_0\underline{1} - \underline{X}\hat{\underline{\beta}}) + 2\sum_{j\in D}\mu_j(\underline{V}_j'\hat{\underline{\beta}}), \tag{3.3}$$

where D denotes the set of latent vectors which are to be deleted from $\hat{\underline{\beta}}$ and the μ_j are Lagrangian multipliers. The resulting estimator is

$$\hat{\underline{\beta}}_{PC} = \sum_{j\in R}\ell_j^{-1}c_j\underline{V}_j, \tag{3.4}$$

where R is the set of latent vectors that are not forced to be orthogonal to $\hat{\underline{\beta}}_{PC}$. The principal component estimator is simply the least squares estimator (3.2) with a chosen set of terms eliminated. Other derivations and motivations for the principal component estimator can be found in Kendall (1957), Massy (1965), and Bock, Yancey, and Judge (1973).

Two criteria are often cited for selecting the terms to eliminate from $\hat{\underline{\beta}}_{PC}$:

(i) delete terms with suitably small ℓ_j

(ii) delete terms which do not sufficiently aid prediction of the response.

The first criterion eliminates terms solely because they are associated with multicollinearities. The second one attempts to assess whether the terms, multicollinear or not, have predictive value, typically utilizing individual F statistics

$$F_j = \ell_j (\underline{V}_j' \hat{\underline{\beta}}_{LS})^2 / \hat{\sigma}^2 \tag{3.5}$$

or by a simultaneous F test

$$F_D = \sum_{j \in D} \ell_j (\underline{V}_j' \hat{\underline{\beta}}_{LS})^2 / n_D \hat{\sigma}^2, \tag{3.6}$$

where the summation in eqn. (3.6) includes all terms which are candidates for deletion (i.e., all those in the set D) and n_D is the number of components deleted. The respective F statistics in eqns. (3.5) and (3.6) are uniformly most powerful for the hypotheses they test, $\underline{V}_j' \underline{\beta} = 0$ and $\underline{V}_j' \underline{\beta} = 0$ $j \in D$, respectively. If the simultaneous test using F_D is chosen, one frequently includes all terms which correspond to multicollinearities that have been identified but it is not necessary to do so or to exclude other terms.

Bock, Yancey, and Judge (1973) studied mean squared error (risk) properties of least squares and preliminary test estimators, of which the

principal component estimator using eqn. (3.6) is a special case. They showed that the mean squared error of the principal component estimator is smaller than that of least squares for small values of

$$\theta = \sum_{j \in D} \ell_j (\underline{V}_j' \underline{\beta})^2 / 2n_D \sigma^2, \tag{3.7}$$

larger for moderate values of θ, and about equal to that of least squares for very large values of θ. Observe that θ is the noncentrality parameter of F_D and large values of θ lead to the retaining of multicollinear terms in $\hat{\underline{\beta}}_{PC}$ through the rejection of the hypothesis that they have no influence on the response.

Clear advantages acrue from the use of $\hat{\underline{\beta}}_{PC}$ with either criterion (i) or (ii) when one can ascertain that $\underline{\beta}$ is orthogonal, or nearly so, to the $\underline{V}_j$ which identify multicollinearities in X (i.e., when θ is small). Either intuition, previous estimates with nonmulticollinear data sets, or theoretical knowledge of the model could lead one to conclude that the coefficient vector is orthogonal to one or more of the multicollinearities. If so, the principal component estimator can be effectively employed, perhaps with the choice of criterion (i) or (ii) depending on one's degree of confidence in the orthogonality of $\underline{\beta}$ to the multicollinearities.

The degree of one's confidence in the orthogonality of $\underline{\beta}$ to the multicollinearities can also be reflected in the choice of significance levels for the F statistics (3.5) and (3.6). The greater the certitude of one's belief in the orthogonality of $\underline{\beta}$ to the multicollinearities, the smaller one should choose the significance level. In other words, the greater one's certitude the stronger the empirical evidence should be to cause one to retain the multicollinear terms. From another argument the same conclusion can be drawn. Since the principal component estimator using either (i) or (ii) as a

criterion has smaller mean squared error than least squares for small values of θ, only large values of θ lead one to desire to retain multicollinear terms. Thus, rejection of H_0 when small significance levels are utilized provides greater assurance that the multicollinear terms should be retained in the estimator regardless of one's possible misgivings about the effects of the multicollinearities or one's belief about the orthogonality of $\underline{\beta}$ to the multicollinear $\underline{V}_j$. If one fails to reject, the multicollinear terms should be deleted.

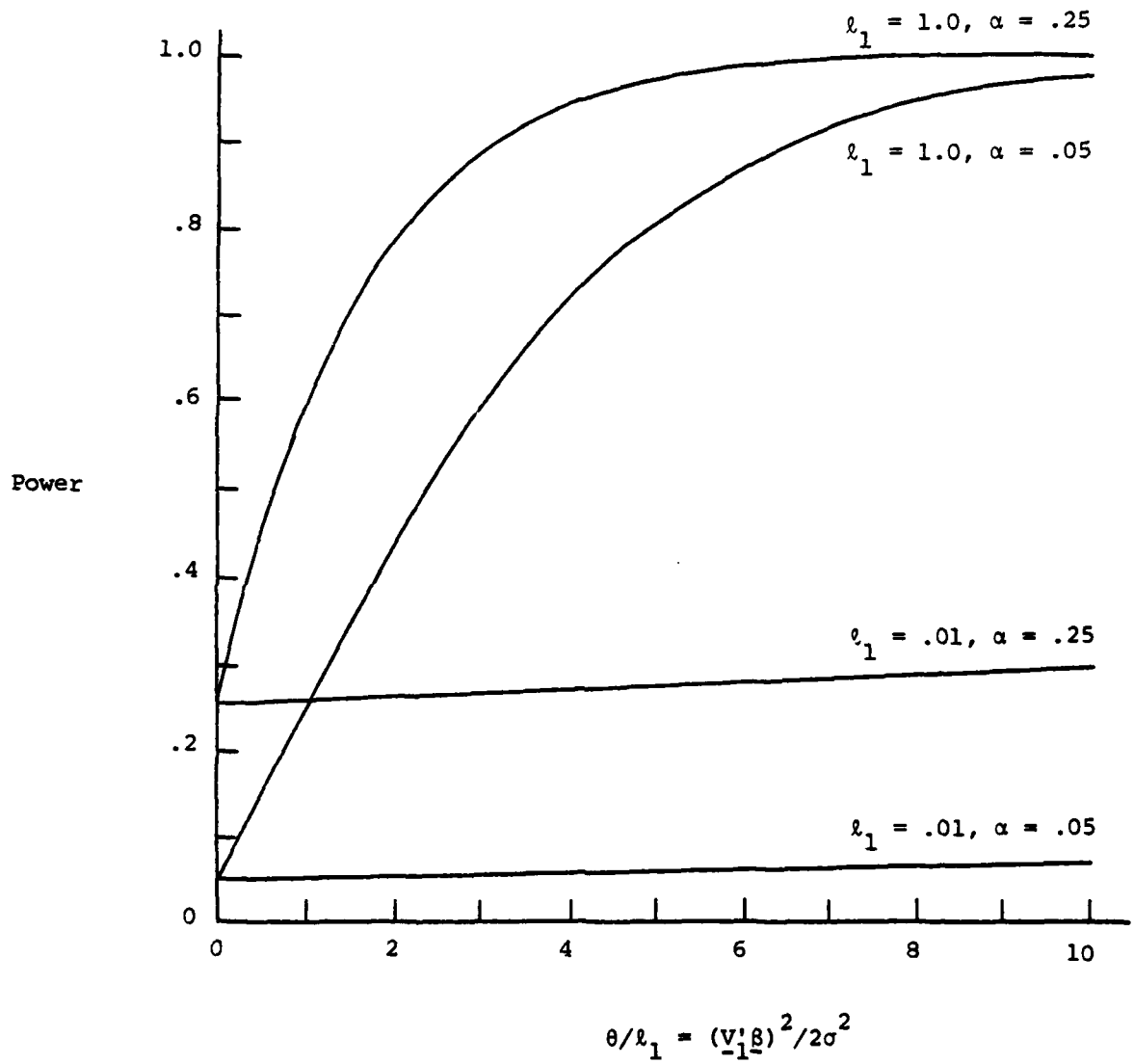
While this strategy is appealing in several respects, additional properties of the F statistics sometimes render inferences unreliable. Since θ can be rewritten in terms of the α_j in eqn. (3.1) as

$$\theta = \sum_{j \in D} \ell_j \alpha_j^2 / 2n_D\sigma^2, \tag{3.8}$$

it is clear that small latent roots can greatly dampen the noncentrality parameter of the F tests over that of an orthogonal X matrix (all $\ell_j = 1.0$). As an illustration of the effect of small latent roots on the power of these tests, Figure 1 graphs power curves for a test of $H_0: \alpha_1 = 0$ vs $H_a: \alpha_1 \neq 0$ for two choices of ℓ_1. The power of the test is greatly reduced when ℓ_1 changes from 1.0 to .01 and can render the test virtually useless for smaller ℓ_1 if α_1 is not very much larger than σ^2. The danger here is that the F test will have poor power because of the multicollinearity (small ℓ_1) and not because α_1 is zero or negligible. Nonrejection of H_0 could cause the elimination of terms from $\hat{\underline{\beta}}_{PC}$ that are necessary for adequate estimation of $\underline{\beta}$. One could compensate for this problem by employing large significance levels but the variance reduction properties of $\hat{\underline{\beta}}_{PC}$ over $\hat{\underline{\beta}}_{LS}$ and the discussions of the previous paragraphs generally dictate that small significance levels be used.

[Insert Figure 1]

FIGURE 1. Power Curves For Testing $H_0: \underline{V}_1'\underline{\beta} = 0$ vs. $H_a: \underline{V}_1'\underline{\beta} \neq 0$ $(n - p - 1 = 10)$.



Consider again the summary information in Table 1. Recall that the pariwise correlation coefficient between P and P^2 is greater than 0.999. With such a large positive correlation, one might expect that the true coefficients should be about equal and the same sign. One very important feature of the pressure readings that was mentioned earlier is that they are measured over a narrow range. The coefficient of variation, $S/\bar{X}$, for the pressure readings is 0.216/29.218 = 0.007. This explains the large positive pairwise correlations between H and H×P (greater than 0.999) and between T and P×T (0.996). One might again, therefore, expect that the true coefficients for the respective pairs of predictor variables (H and H×P, T and P×T) should be of the same sign and about equal in magnitude. Thus the true coefficient vector $\underline{\beta}$ should be orthogonal to both $\underline{V}_1$ and $\underline{V}_2$, a property not reflected in the least squares estimates because of the great influence of $\underline{V}_1$ and $\underline{V}_2$ on $\hat{\underline{\beta}}_{LS}$.

In contrast to the least squares estimates are the principal component estimates, displayed in Table 2. First consider whether to delete the two latent vectors discussed above. The arguments posed in the last paragraph could lead one to eliminate $\underline{V}_1$ and $\underline{V}_2$ without performing a preliminary test. If one was uncertain as to the correctness of those arguments, the summary information given in Table 2 on F_1, F_2, and F_D (deleting components 1, 2) all suggest the same result: delete both. Now observe that when $\underline{V}_1$ is deleted, the signs on H, H×P, and H×T are all reversed, with H and H×T becoming negative. The magnitudes on many of the estimates are reduced as well. This trend continues when $\underline{V}_2$ is eliminated. It is reassuring to observe that when $\underline{V}_1$ and $\underline{V}_2$ are deleted humidity has the largest coefficient and its sign is negative, but the relative magnitudes of temperature, P×T, H×T, and P^2 as well as the signs on H×P and P^2 are still disturbing (H×P should have

the same sign as H, and P^2 should have the same sign as P). It is now time to examine the remaining latent roots and latent vectors of X'X.

[Insert Table 2]

Table 3 contains four additional latent roots and the corresponding latent vectors of X'X. Carefully examining $\underline{V}_3$, one observes that the elements for H and P^2 are about equal in magnitude and of the same sign (reflecting the negative correlation between H and P and the equivalence of P and P^2), those of P and H×P are about equal in magnitude and of the same sign (again indicating the negative correlation between P and H since H is highly correlated with H×P), and the elements corresponding to T and P×T are about equal and opposite in sign (showing the correlation between T and P×T). One can again argue that $\underline{\beta}$ is orthogonal to $\underline{V}_3$ or use F tests (see Table 2) to conclude that $\underline{V}_3$ should be deleted. When $\underline{V}_1$, $\underline{V}_2$, and $\underline{V}_3$ are deleted from $\hat{\underline{\beta}}_{PC}$, the signs and magnitudes of the estimated coefficients appear to be consistent with the investigators' a priori knowledge, except perhaps that the magnitude of temperature is too large. Note, however, the large drops in magnitudes when $\underline{V}_3$ is deleted.

[Insert Table 3]

The remaining three latent vectors displayed in Table 3 can be assessed similarly to the above three. One of these vectors, however, deserves special note. The sixth latent vector corresponds to a latent root which is five orders of magnitude larger than the smallest one. Its elimination will not result in as substantial a reduction in variance as any of the previous latent vectors. The large elements in $\underline{V}_6$ correspond to H, H×P, and H^2, all of which involve humidity. Since humidity is believed to be an important predictor of NO_x emissions, deletion of this latent vector could result in

Table 2. Principal Component Estimates for NO_x Emissions Data, Quadratic Fit

Predictor Variable	Components Deleted						
	None	$\underline{v}_1$	$\underline{v}_1$, $\underline{v}_2$	$\underline{v}_1$ to $\underline{v}_3$	$\underline{v}_1$ to $\underline{v}_4$	$\underline{v}_1$ to $\underline{v}_5$	$\underline{v}_1$ to $\underline{v}_6$
Humidity (H)	44.738	-13.547	-11.520	-1.178	-1.372	-.561	-.060
Pressure (P)	69.736	20.371	9.967	.175	.136	.068	.224
Temperature (T)	.734	14.551	-6.754	-.887	-.248	-.006	.007
P×T	-1.710	-15.973	4.962	-1.082	-.530	-.007	.027
H×P	-46.019	12.525	8.598	-1.165	-1.358	-.600	-.059
H×T	.797	-1.059	1.506	1.337	1.744	-.296	-.048
P^2	-69.005	-19.036	-10.374	.101	.081	.083	.224
T^2	.971	1.924	1.581	1.878	.627	.077	.007
H^2	.458	1.408	1.008	.701	.706	1.047	-.006
R^2	.835	.823	.818	.808	.805	.775	.690
$\hat{\sigma}^2$	.00241	.00252	.00251	.00258	.00256	.00287	.00385

Component Deleted	1	2	3	4	5	6
λ_j	$.517\times10^{-6}$	$.196\times10^{-5}$	$.109\times10^{-4}$	$.647\times10^{-3}$	$.239\times10^{-2}$	$.239\times10^{-1}$
F_j	2.612	.897	2.168	.677	6.108	17.556
Significance Probability	.10<p<.25	.25<p	.10<p<.25	.25<p	.01<p<.025	p<.001

Components Deleted	1,2	1,2,3	1,2,3,4	1,2,3,4,5	1,2,3,4,5,6
F_D	1.754	1.892	1.588	2.492	5.003
Significance Probability	.10<p<.25	.10<p<.25	.10<p<.25	.10<p<.25	p ≃ .001

Table 3. Additonal Latent Vectors of X'X, Quadratic Model

Predictor Variable	$\ell_3 = .109\times10^{-4}$ $\underline{V}_3$	$\ell_4 = .647\times10^{-3}$ $\underline{V}_4$	$\ell_5 = .239\times10^{-2}$ $\underline{V}_5$	$\ell_6 = .239\times10^{-1}$ $\underline{V}_6$
Humidity (H)	-.473	.123	-.327	-.378
Pressure (P)	.447	.024	.027	-.117
Temperature (T)	-.268	-.402	-.098	-.010
P×T	.276	-.347	-.211	-.025
H×P	.446	.121	-.305	-.407
H×T	.008	-.257	.823	-.187
P^2	-.479	.013	-.001	-.107
T^2	-.014	.788	.221	.053
H^2	.014	-.003	-.138	.793

large bias. Coupled with the relatively smaller reduction in variance from deleting $\underset{\sim}{V}_6$ relative to the other latent vectors, the potential for inducing bias suggests that it should be examined carefully before allowing its removal. The F statistic in Table 2 confirms that this component is significant at extremely small significance levels and should be retained.

Based on these considerations, the principal component estimates ultimately selected consisted of the elimination of $\underset{\sim}{V}_1$ through $\underset{\sim}{V}_5$. The standardized estimates shown in column seven of Table 2 have relatively large magnitudes on all the humidity terms, although the magnitudes are greatly reduced from the least squares quadratic fit and are of the same relative size as those in the linear fit. The signs on the humidity terms are also consistent with the investigators' belief that humidity should have a negative effect on NO_x. The other variables, perhaps with the exception of T and P×T, are seen to have a lesser but nonnegligible effect on NO_x.

From the detailed nature of the foregoing analysis, it should be apparent that biased estimation is neither routine nor automatic. The desire of some data analysts for an automated analysis is both a source of criticism of biased estimation and an invitation to poor results. On the other hand, careful and detailed examination of the data base, and in particular multicollinearities, leads to a better understanding of the resulting estimates and more fruitful, consistent conclusions.

3.2 Latent Root Regression

Latent root regression was spawned by a desire to more effectively assess multicollinearities. Rather than examining only the latent roots and latent vectors of X'X, latent root regression also investigates the latent roots and latent vectors of A'A, the $(p + 1) \times (p + 1)$ matrix of correlations of response

and predictor variables; i.e., let $A = [\underline{Y}^*:X]$ and $Y_i^* = \eta^{-1}(Y_i - \bar{Y})$, where $\eta^2 = \Sigma(Y_i - \bar{Y})^2$. Multicollinearities are first identified by studying the pairwise correlations in X'X, the variance inflation factors (Marquardt (1970), Marquardt and Snee (1975)), and the latent roots and latent vectors of X'X. Multicollinearities are then assessed by analyzing the latent roots and latent vectors of A'A.

Let $0 \leq \lambda_0 \leq \lambda_1 \leq \ldots \leq \lambda_p$ denote the latent roots of A'A and $\underline{\gamma}_0$, $\underline{\gamma}_1$, ..., $\underline{\gamma}_p$ the corresponding latent vectors. Partition $\underline{\gamma}_j$ as $\underline{\gamma}_j' = (\gamma_{0j}, \underline{\delta}_j')$ so that $\underline{\delta}_j$ contains the last p elements of $\underline{\gamma}_j$. If λ_j is sufficiently close to zero, $\underline{\gamma}_j$ identifies a multicollinearity in A just as $\ell_j \approx 0$ points out a multicollinearity of X that is defined by $\underline{V}_j$. If, in addition to λ_j sufficiently close to zero, the magnitude of γ_{0j} is also close to zero, $\underline{\gamma}_j$ identifies a multicollinearity which involves only the predictor variables and not the response, a "nonpredictive" multicollinearity (see Webster, Gunst, and Mason (1974) and White and Gunst (1979)). Latent root regression eliminates nonpredictive multicollinearities from the resulting estimator.

In terms of the $\underline{\delta}_j$, the latent root estimator is given by

$$\hat{\underline{\beta}}_{LR} = \sum_{j \in R} \lambda_j^{-1} f_j \underline{\delta}_j, \quad \text{where } f_j = -\eta\gamma_{0j} / \sum_{i \in R} \gamma_{0i}^2 \lambda_i^{-1}. \tag{3.9}$$

In eqn. (3.9) as in the formula for $\hat{\underline{\beta}}_{PC}$, R refers to the set of indices of latent vectors that are retained in the latent root estimator; i.e., all latent vectors except those which identify nonpredictive multicollinearities. The latent root estimator minimizes

$$\phi = (Y - \hat{\beta}_0\underline{1} - X\hat{\underline{\beta}})'(\underline{Y} - \hat{\beta}_0\underline{1} - X\hat{\underline{\beta}}) + 2 \sum_{j \in D} \mu_j (\underline{\delta}_j'\hat{\underline{\beta}} - \eta\gamma_{0j}), \tag{3.10}$$

which is similar in form to eqn. (3.3), especially since γ_{0j} is close to zero for nonpredictive multicollinearities. If the multicollinearities of A'A

are different from those of X'X, however, the latent root estimator can be quite different from both the least squares and the principal component estimator.

As expressed in eqn. (3.9) the latent root estimator is a complicated function of the response variable and its distributional properties are currently unknown. Apart from the intuitive appeal of being able to assess the predictiveness of multicollinearities and adjust for nonpredictive ones, the efficacy of $\hat{\underline{\beta}}_{LR}$ has been argued primarily from simulation studies (see Gunst, Webster and Mason (1976) and Gunst and Mason (1977a)), as is the case for many other biased estimators. Recently, White and Gunst (1979) developed asymptotic inference procedures which can be utilized to refine the assessment of multicollinearities when sample sizes are large.

One facet of latent root regression suggests that it is a valuable alternative to principal component regression. Not only can multicollinearities in X'X be assessed for their predictive value, occasionally new multicollinearities appear in A'A which are different from those of X'X. Necessarily these multicollinearities involve the response variable and generally have predictive value. They can take on at least two forms: "crossovers" and "distortions."

A crossover occurs if one of the latent vectors of X'X, say $\underline{V}_k$, displaces another one, say $\underline{V}_m$, in A'A when $\ell_m < \ell_k$. For example if $\underline{V}_1$ corresponds to the smallest latent root ℓ_1 of X'X and $\underline{V}_2$ corresponds to the next smallest one ℓ_2, yet in A'A $\underline{\delta}_0 \simeq \underline{V}_2$ and $\underline{\delta}_1 \simeq \underline{V}_1$, a crossover is said to have occurred ("crossover" since now $\underline{V}_2 \simeq \underline{\delta}_0$ corresponds to the smallest latent root of A'A and $\underline{V}_1 \simeq \underline{\delta}_1$ to the second smallest). Regardless of the apparent magnitude of γ_{00}, the elements of $\underline{\gamma}_0$ identify a predictive multicollinearity. This is

because if $\lambda_0 < \ell_1$ and γ_{00} is actually zero,

$$\lambda_0 = \underline{\gamma}_0'A'A\underline{\gamma}_0 = \underline{\delta}_0'X'X\underline{\delta}_0 < \underline{V}_1'X'X\underline{V}_1 = \ell_1.$$

But this is a contradiction since $\underline{V}_1$ is the unique unit length vector which minimizes the quadratic form $\underline{u}'X'X\underline{u}$. The reason for the caution about the apparent magnitude of γ_{00} is that if $\underline{\beta}$ is a linear combination of only multicollinear latent vectors of X'X (those associated with small latent roots), the first element of $\underline{\gamma}_0$ can be small even if $\underline{\gamma}_0$ identifies a predictive multicollinearity (see White and Gunst (1979)).

Similarly, if $\underline{\beta}$ is a linear combination of some multicollinear latent vectors of X'X and some nonmulticollinear ones, the multicollinearities in A'A can be distortions of those of X'X. For example, suppose both α_1 and α_p are large in eqn. (3.1) and all other α_j are zero. If σ^2 is sufficiently small, $\underline{\delta}_0$ will be approximately proportional to $\alpha_1\underline{V}_1 + \alpha_p\underline{V}_p$ and none of the other $\underline{\delta}_j$ will be approximately equal to $\underline{V}_1$. Thus $\underline{V}_1$ is "distorted" in the latent vectors of A'A although the other multicollinear latent vectors of X'X might not be.

The value of this information is that one need not restrict inferences on multicollinearities to the individual latent vectors of X'X or to a linear combination of only the multicollinear ones, as in eqns. (3.5) and (3.6). If $\underline{\beta}$ is actually equal to one of the $\underline{V}_j$ or a linear combination of only the multicollinear ones, analysis of predictive multicollinearities using the latent roots and latent vectors of A'A cannot have greater statistical power than the appropriate F statistics. However, an analysis of the latent roots and vectors of A'A can potentially detect whether $\underline{\beta}$ only partially involves the latent vectors associated with multicollinearities.

Turning now to an analysis of the emissions data, Table 4 lists the six smallest latent roots of A'A and their associated latent vectors. Comparing these latent roots and latent vectors with those of X'X, neither crossovers nor distortions are apparent. All six latent vectors are associated with small latent roots but the first element of $\underset{\sim}{\gamma}_5$ is clearly not close enough to zero to be judged a nonpredictive multicollinearity. This finding supports the conclusions drawn above on $\underset{\sim}{v}_6$ from the principal component analysis and is again due to the large weight on H^2 in the latent vector as well as the moderately large weights on H, H×P, and H×T.

[Insert Table 4]

The very small γ_{0j} values for the first four latent vectors in Table 4 and the discussions in the previous section of the multicollinearities lead immediately to the elimination of these latent vectors from $\hat{\underset{\sim}{\beta}}_{LR}$. Whether $\underset{\sim}{\gamma}_4$ should also be eliminated is questionable. Its first element is larger than any of the previous four but the decision of whether it is sufficiently close to zero to be labeled nonpredictive is unclear. Since the three elements of $\underset{\sim}{\gamma}_4$ that have large weights all involve humidity, one might prefer to leave this vector in the estimator. On the other hand, one could argue as follows that the multicollinearity is theoretically nonpredictive. As with $\underset{\sim}{\gamma}_3$ and $\underset{\sim}{\gamma}_5$ (as well as $\underset{\sim}{v}_4$, $\underset{\sim}{v}_5$, and $\underset{\sim}{v}_6$), $\underset{\sim}{\gamma}_4$ is a three-variable multicollinearity which is actually reflecting two two-variable ones. The correlation between H and H×T is .992 and that between H×P and H×T is .992, so that $X_1 \approx X_6$ and $X_5 \approx X_6$. Implied by these multicollinearities is that $X_1 + X_5 \approx 2X_6$ or $2X_6 - X_1 - X_5 \approx 0$. Observe that the magnitude of the element of $\underset{\sim}{\gamma}_4$ corresponding to H×T is roughly twice the size and opposite in sign of that of H and H×T, just as the relationship $2X_6 - X_1 - X_5$ indicates it should be. Now due to the large pairwise correlations mentioned above,

Table 4. Six Latent Roots and Latent Vectors of A'A, Quadratic Fit

Predictor Variable	$\lambda_0 = .480\times10^{-6}$ $\underline{Y}_0$	$\lambda_1 = .191\times10^{-5}$ $\underline{Y}_1$	$\lambda_2 = .103\times10^{-4}$ $\underline{Y}_2$	$\lambda_3 = .635\times10^{-3}$ $\underline{Y}_3$	$\lambda_4 = .206\times10^{-2}$ $\underline{Y}_4$	$\lambda_5 = .170\times10^{-1}$ $\underline{Y}_5$
Response	.000	.001	-.002	-.008	-.039	-.151
Humidity (H)	-.519	.086	-.482	.105	-.361	-.356
Pressure (P)	-.463	-.309	.435	.023	.025	-.071
Temperature (T)	.111	-.639	-.281	-.405	-.081	.006
P×T	-.115	.630	.284	-.356	-.198	.003
H×P	.521	-.142	.448	.101	-.347	-.388
H×T	-.016	.077	.012	-.228	.812	-.261
P^2	.468	.257	-.467	.017	.007	-.054
T^2	.007	-.012	-.009	.797	.194	-.024
H^2	.008	-.013	.016	.004	-.074	.790

one might expect that $\beta_6 \simeq \beta_1$ and $\beta_6 \simeq \beta_5$ so that $\underline{\beta}'\underline{\delta}_4 \simeq 0$. Thus one could argue for the elimination of $\underline{\gamma}_4$ as well as $\underline{\gamma}_0$ through $\underline{\gamma}_3$.

Table 5 displays the latent root estimates of $\underline{\beta}$ as the first six latent vectors are sequentially deleted from $\hat{\underline{\beta}}_{LR}$. In the fifth set of estimates, $\underline{\gamma}_0$ through $\underline{\gamma}_4$ have been removed. All four humidity predictor variables have large coefficient estimates with $\hat{\beta}_6 \simeq \hat{\beta}_1$ and $\hat{\beta}_6 \simeq \hat{\beta}_5$. The estimates are not inconsistent with the investigators' a priori knowledge nor are they substantially different from the principal component estimates. This is not surprising since neither crossovers nor distortions occurred in this data set among the latent vectors of X'X and A'A that identify multicollinearities.

[Insert Table 5]

A major defect of the latent root estimator is the lack of exact distributional theory. One must rely on a careful examination of the data base and diagnostic statistics that are associated with the methodology in order to draw adequate inferences with latent root regression. Even if the asymptotic theoretical properties can be invoked to assist in drawing inferences, a careful examination of X'X, its latent roots and vectors, the latent roots and vectors of A'A, etc. is indispensable and cannot be automated. Compensating for the added effort, however, is a better comprehension of the data base and its limitations and a far more informed understanding of the regression estimates, including their signs and magnitudes.

3.3 Ridge Regression

Perhaps the single most influential catalyst to the current widespread popularity of biased regression estimation is Hoerl and Kennard's (1970a, b) development of ridge regression. By simply adding a small positive quantity k to the diagonal elements of X'X, an entire family of estimators can be generated:

Table 5. Latent Root Estimates for NO_x Emissions Data, Quadratic Fit

Predictor Variable	Components Deleted						
	None	$\underline{Y}_0$	$\underline{Y}_0$, $\underline{Y}_1$	$\underline{Y}_0$ to $\underline{Y}_2$	$\underline{Y}_0$ to $\underline{Y}_3$	$\underline{Y}_0$ to $\underline{Y}_4$	$\underline{Y}_0$ to $\underline{Y}_5$
Humidity (H)	44.738	-14.127	-11.573	-1.167	-1.373	-.509	-.003
Pressure (P)	69.736	19.504	9.604	.157	.119	.063	.237
Temperature (T)	.734	14.186	-6.935	-.883	-.195	.017	.013
P×T	-1.710	-15.624	5.136	-1.085	-.487	.030	.036
H×P	-46.019	13.060	8.632	-1.175	-1.374	-.553	.002
H×T	.797	-1.015	1.532	1.337	1.761	-.401	-.042
P^2	-69.005	-18.204	-10.026	.120	.093	.088	.237
T^2	.971	1.927	1.581	1.877	.527	.028	-.008
H^2	.458	1.407	1.003	.701	.708	1.049	-.111
R^2	.835	.822	.818	.808	.804	.771	.671
$\hat{\sigma}^2$	.00241	.00252	.00251	.00258	.00256	.00291	.00409

$$\hat{\underline{\beta}}_{RR} = (X'X + kI)^{-1}X'\underline{Y}$$

$$= \sum_{j=1}^{p} (\ell_j + k)^{-1} c_j \underline{V}_j \quad . \tag{3.11}$$

These estimators include least squares ($k = 0$) and the null vector ($k = \infty$) as extremes. From eqn. (3.11) one can see that, unlike principal component and latent root regression analyses, the ridge estimator deletes none of the latent vectors of X'X from the analysis but decreases the weights (from least squares) on each of them. In this manner the effects of multicollinearities are lessened.

Several important theoretical results are provided by Hoerl and Kennard (1970a). First, the ridge estimator can be derived by minimizing the residual sum of squares subject to a fixed value of $\hat{\underline{\beta}}'\hat{\underline{\beta}}$; i.e., $\hat{\underline{\beta}}_{RR}$ minimizes

$$\phi = (\underline{Y} - \hat{\beta}_0\underline{1} - X\hat{\underline{\beta}})'(Y - \hat{\beta}_0 1 - X\hat{\underline{\beta}}) + \mu(\hat{\underline{\beta}}'\hat{\underline{\beta}} - r). \tag{3.12}$$

In the minimization of (3.12), the lagrangian multiplier μ becomes k in $\hat{\underline{\beta}}_{RR}$. An equivalent derivation of $\hat{\underline{\beta}}_{RR}$ minimizes the length of the coefficient vector subject to a fixed increase in the residual sum of squares over that of least squares; i.e., $\hat{\underline{\beta}}_{RR}$ minimizes

$$\phi = \hat{\underline{\beta}}'\hat{\underline{\beta}} + \mu[(\hat{\underline{\beta}} - \hat{\underline{\beta}}_{LS})'X'X(\hat{\underline{\beta}} - \hat{\underline{\beta}}_{LS}) - s]. \tag{3.13}$$

Again, $\hat{\underline{\beta}}_{RR}$ minimizes eqn. (3.13) with $\mu = k^{-1}$. Both of these properties of ridge estimators are useful characterizations; moreover, if the researcher can specify values for r in eqn. (3.12) or s in eqn. (3.13), unique solutions for k can be obtained.

Often considered the most important - and most controversial - rationale for ridge regression is the "existence theorem" proven by Hoerl and Kennard (1970a). The authors show that there always exists a range of values of the

ridge parameter (k) for which $mse(\hat{\underline{\beta}}_{RR}) < mse(\hat{\underline{\beta}}_{LS})$, where

$$mse(\hat{\underline{\beta}}) = E[(\hat{\underline{\beta}} - \underline{\beta})'(\hat{\underline{\beta}} - \underline{\beta})]. \tag{3.14}$$

Two recurring criticisms of the existence theorem are (i) the range on k depends on the unknown model parameters, and (ii) the criterion (3.14) is only one of many important criteria for assessing regression estimators (e.g., $E[(\hat{\underline{\beta}} - \underline{\beta})'X'X(\hat{\underline{\beta}} - \underline{\beta})]$ is another). Progress has been made in answering criticism (ii); e.g., Theobald (1974) generalized the existence theorem to include any criteria of the form

$$mse(\hat{\underline{\beta}}) = E[(\hat{\underline{\beta}} - \underline{\beta})'M(\hat{\underline{\beta}} - \underline{\beta})],$$

where M is at least positive semidefinite.

No completely satisfactory response has been forthcoming to the first criticism. Indeed, further controversy has emanated from a related problem, procedures for selecting the ridge parameter. Most of the more popular choices of k rely on stochastic techniques (ridge trace, estimation of the upper bound suggested by the existence theorem, etc.) and thereby invalidate the application of the existence theorem. No ridge estimator has yet been devised which can guarantee a smaller mean squared error than least squares for all model configurations.

Admission of this last point does not, however, negate the advantages of ridge regression when predictor variables are multicollinear. Least squares parameter estimates such as those presented in Table 1 for the interaction and quadratic fits are clearly unacceptable. The reasons for the patterns in the signs and magnitudes have been traced directly to the multicollinearities in X and cannot be attributable to the true functional relationships between response and predictor variables. While it is true that ridge regression cannot guarantee a smaller mean squared error than least squares, especially for

this severely multicollinear a data set it is also true that least squares cannot guarantee a smaller mean squared error than ridge regression. Ridge estimates might, however, produce far more reasonable coefficient estimates with no essential increase in the residual sum of squares. Let us expand on this statement.

The residual sum of squares for ridge estimators can be written as

$$\begin{aligned} SSE_{RR} &= (\underline{Y} - \hat{\beta}_0\underline{1} - X\underline{\hat{\beta}}_{RR})'(\underline{Y} - \hat{\beta}_0\underline{1} - X\underline{\hat{\beta}}_{RR}) \\ &= (\underline{Y} - \hat{\beta}_0\underline{1} - X\underline{\hat{\beta}}_{LS})'(\underline{Y} - \hat{\beta}_0\underline{1} - X\underline{\hat{\beta}}_{LS}) + (\underline{\hat{\beta}}_{RR} - \underline{\hat{\beta}}_{LS})'X'X(\underline{\hat{\beta}}_{RR} - \underline{\hat{\beta}}_{LS}) \\ &= SSE_{LS} + \omega(\underline{\hat{\beta}}_{RR}), \end{aligned}$$

where $\omega(\underline{\hat{\beta}}_{RR})$ is the increase in the residual sum of squares over that of least squares. An alternative expression for $\omega(\underline{\hat{\beta}}_{RR})$ is

$$\omega(\underline{\hat{\beta}}_{RR}) = \sum_{j=1}^{p} \ell_j [\underline{V}_j'(\underline{\hat{\beta}}_{RR} - \underline{\hat{\beta}}_{LS})]^2. \tag{3.15}$$

If k is chosen suitably small, eqn. (3.11) shows that only the weights on latent vectors of X'X corresponding to very small latent roots will be substantially altered from those of least squares. Those latent vectors, however, receive the smallest weights in eqn. (3.15). Thus the ridge estimator can cause $\underline{\hat{\beta}}_{RR}$ and $\underline{\hat{\beta}}_{LS}$ to differ greatly in dimensions corresponding to multicollinearities in X while not appreciably increasing the residual sum of squares over that of least squares. This is further evidence of the tenuous nature of least squares estimates when predictor variables are multicollinear: there can be a number of estimates which are numerically quite different from least squares but whose residual sums of squares are for all practical purposes just as small. This notion of directions in estimation space for which coefficient estimates can change radically without markedly increasing the residual sum of squares is the same principle which underlies "ridge

analysis" of response surfaces and was one of the first motivations of ridge regression (see Hoerl (1962)).

Table 6 displays ridge estimates for several choices of the ridge parameter, k. All of the sets of estimates shown in the table tend to have magnitudes that are greatly reduced from least squares but which are of the same order as the least squares fit to only the linear terms. In all the ridge estimates H, H×P, and H^2 have relatively large magnitudes with H and H×P having negative signs. Some of the temperature variables have moderate to large magnitudes for small k and some of the pressure variables have moderate to large ones for the larger k values.

[Insert Table 6]

The ridge estimates in Table 6 represent a continuous damping of the multicollinearities rather than the discrete inclusion or exclusion of each as with principal component and latent root regressions. The ridge estimates thereby represent a compromise among various principal component or latent root estimates. With this data set even the small value of the ridge parameter, k = .005, results in a ridge estimate which is compromise between the deletion of four or five multicollinearities; i.e., even this small k value drastically reduces the weights associated with the first four or five latent vectors of X'X. Comparing the ridge estimates for k = .005 with the principal component (Table 2) and latent root (Table 5) estimates, except for P and P^2 each of the individual ridge estimates lies between the values of the corresponding principal component or latent root estimates for four and five deleted components. For example, the ridge estimate for the coefficient of H, -.733, is between -1.372 and -.561 for $\hat{\beta}_{PC}$ and between -1.373 and -.509 for $\hat{\beta}_{LR}$. Even the ridge estimated coefficients for P and P^2 are very close to the corresponding ones for the other two biased estimators.

Table 6. Ridge Estimates for NO_x Emissions Data, Quadratic Fit

Predictor Variable	Ridge Parameters				
	k = 0	k = .005	k = .01	k = .05	k = .10
Humidity (H)	44.738	-.733	-.568	-.262	-.181
Pressure (P)	69.736	.151	.143	.169	.176
Temperature (T)	.734	-.161	-.090	-.016	-.003
P×T	-1.710	-.227	-.128	-.014	.005
H×P	-46.019	-.713	-.567	-.268	-.185
H×T	.797	.359	.145	-.044	-.057
P^2	-69.005	.076	.107	.164	.174
T^2	.971	.386	.239	.072	.044
H^2	.458	.754	.669	.312	.176
R^2	.835	.789	.778	.738	.719
$\hat{\sigma}^2$	.00241	.00308	.00324	.00383	.00410

When k is increased to .01, the ridge estimates become closer to the previous two biased estimates in which five components are deleted. When k = .05, the ridge estimates whose magnitudes are greater than .10 all lie between those of $\hat{\beta}_{PC}$ and $\hat{\beta}_{LR}$ for five and six components deleted, as do all the ridge estimates for k = .10. Numerically these trends reinforce the theoretically apparent relationship shown in eqn. (3.11): as k increases, the effects on $\hat{\beta}_{RR}$ of latent vectors corresponding to small latent roots are lessened.

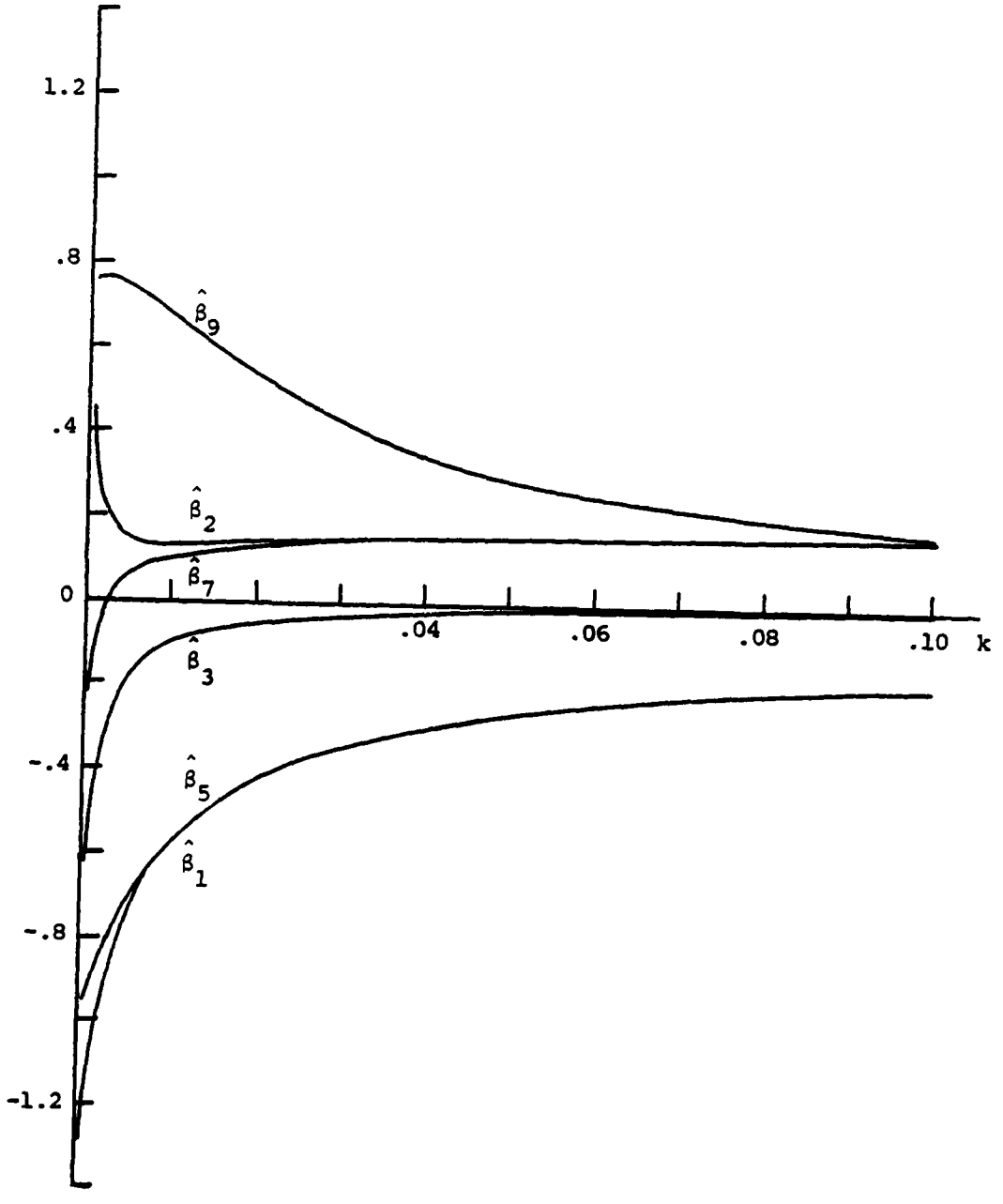
Which value of the ridge parameter should be used? This is perhaps the single most controversial question surrounding the application of ridge regression. Figure 2 contains a ridge trace, plots of the estimated coefficients as a function of k, over the range $.0005 \leq k \leq .10$. For ease of viewing, only six of the ridge estimates are plotted, the estimates for H, P, T, H×P, P^2, and H^2. The first point plotted corresponds to k = .0005. Even with this small a value of the ridge parameter two of the estimates differ in sign with the least squares estimates ($\hat{\beta}_1$ and $\hat{\beta}_3$). As k increases $\hat{\beta}_7$ also changes sign. The magnitudes of the estimates, moreover, are much smaller than the least squares estimates over the entire range of k displayed in the figure.

[Insert Figure 2]

Choosing k by the ridge trace method is recognized to be highly subjective. Following the advice of Hoerl and Kennard (1970a), the trace appears to have stabilized in the range $.005 < k < .01$. Either of the first two sets of ridge estimates in Table 6, or some intermediate choice, could be viewed as a reasonable choice.

Apart from one's determination of where the trace stabilizes, other problems affect ridge traces. Figure 3 displays three sets of ridge traces

FIGURE 2. Ridge Trace For Selected Parameter Estimates.



for two of the coefficient estimates, $\hat{\beta}_1$ and $\hat{\beta}_9$. The three curves for each estimate differ only with respect to the range of k that is plotted: $.01 \leq k \leq 1.0$ ($\rho = 1.0$), $.001 \leq k \leq .10$ ($\rho = .10$), and $.0001 \leq k \leq .01$ ($\rho = .01$). Decisions regarding when the traces stabilize appear to be dependent on which range of k one plots.

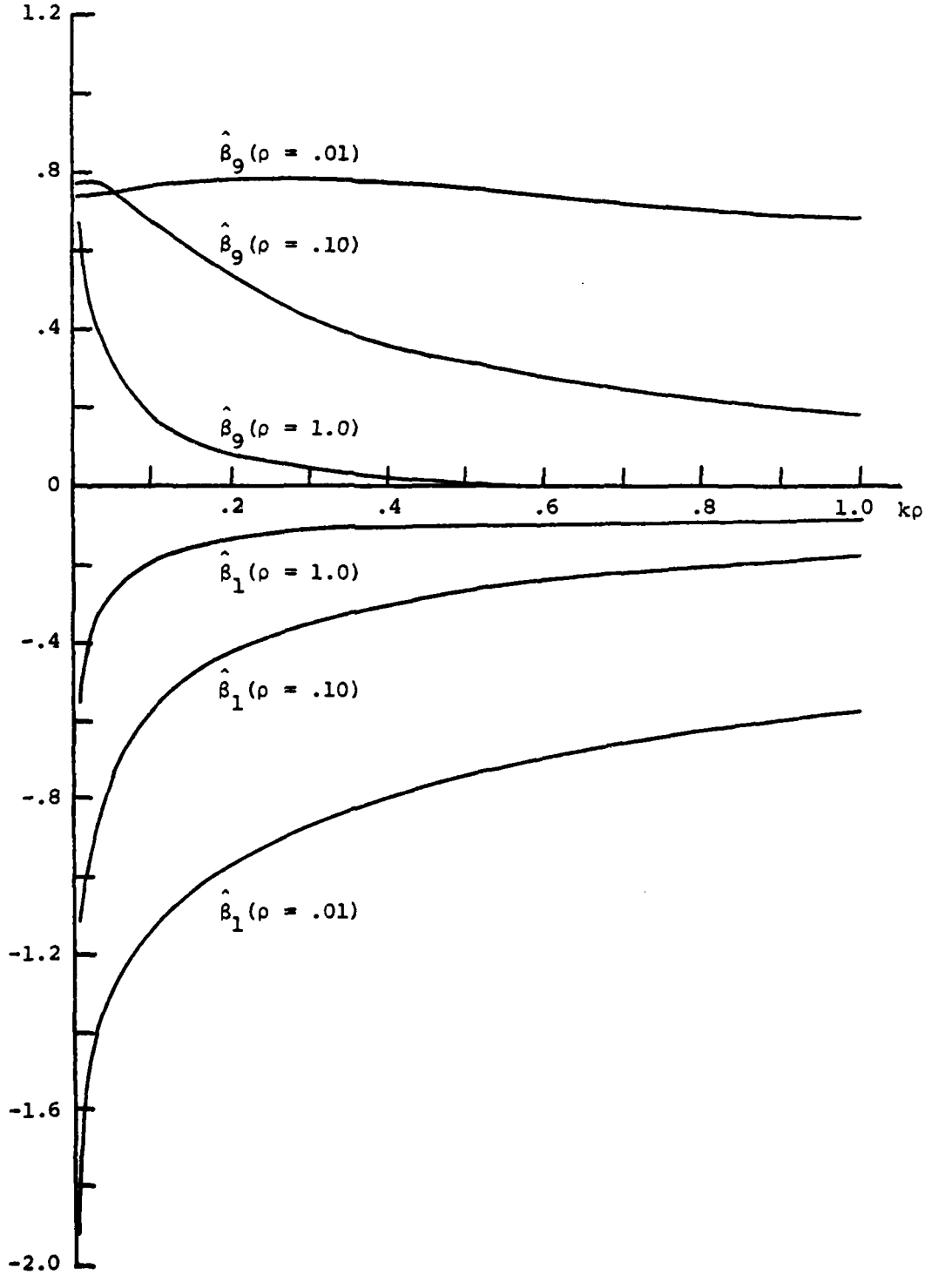
[Insert Figure 3]

Stochastic and nonstochastic rules for selecting ridge parameter values have frequently occurred in the statistical literature over the last ten years (e.g., Dempster, Schatzoff, and Wermuth (1977), Hoerl, Kennard, and Baldwin (1975), and McDonald and Galarneau (1975)). Numerous as these rules are, they also tend to suggest different choices of the ridge parameter. For example, Hoerl, Kennard, and Baldwin (1975) recommended a stochastic estimator of k which has also performed reasonably well in other simulations. (e.g. Gibbons (1978), Gunst and Mason (1977a)). For this data set the estimate of k is

$$\hat{k} = p\hat{\sigma}^2/\hat{\underline{\beta}}'_{LS}\hat{\underline{\beta}}_{LS} = .000002.$$

A nonstochastic rule for selecting k which has been advocated is to choose k so that the largest variance inflation factor is less than some convenient value, say 10. For the emissions data the resulting value of the ridge parameter is k = .02. Whether one's preference is to use either of these values or one from a ridge trace depends on one's familiarity and experience with the various procedures. For many data sets in which the multicollinearities are not as numerous or severe as this one, all these techniques can yield ridge estimates that are similar and the choice of a specific rule is not critical. For this data set - as with the choice of a specific principal component or latent root estimate - the choice is important, as the differences in the estimates in Table 7 attest.

FIGURE 3. Ridge Traces For Three Ranges of k.



[Insert Table 7]

Although the foregoing discussion seems to lead to a morass of confusion, it is intended to stress a point: ridge regression cannot be mechanically applied without regard to characteristics of the data set being analyzed. This warning was stressed with reference to biased estimators in general in Section 2 and with special reference to principal component and latent root regression earlier in this section, but it cannot be overemphasized. Thus while the stochastic rule proposed by Hoerl, Kennard, and Baldwin has performed well in their simulation and others, it is not recommended for this data set. Since the emissions data is so highly multicollinear and the least squares estimates are thereby inflated, the denominator of $\hat{k}$ is much larger than it should be and $\hat{k}$ is extremely close to zero. The corresponding ridge estimates still have magnitudes and signs that are inconsistent with the investigators' a priori suspicions.

Use of one of the values suggested by the ridge trace in Figure 2, $k = .005$, yields ridge estimates that are more reasonable than the stochastic estimate of k. Even so, one could question the relative magnitudes of the non-humidity coefficients and the sign on H×T in light of the suspected dominance of humidity and the correlations in X'X. Use of the other value suggested by the ridge trace, $k = .01$ (see Table 6), or the value suggested by the maximum variance inflation factor criterion, $k = .02$, lessens these last concerns. All three of these ridge estimates (viz., $k = .005, .01, .02$) are similar and the choice should be based on external information insofar as possible. Here, however, the final choice is not as critical as it was between least squares, the ridge estimates based on $\hat{k}$, and one of these three.

Table 7. Ridge Estimates for Several Selection Rules

Predictor Variable	Selection of the Ridge Parameter: Stochastic Rule k = .000002	Ridge Trace k = .005	Maximum VIF k = .02
Humidity (H)	1.058	-.733	-.420
Pressure (P)	23.740	.151	.151
Temperature (T)	1.858	-.161	-.047
P×T	-3.398	-.227	-.063
H×P	-3.004	-.713	-.426
H×T	.592	.359	.020
P^2	-23.304	.076	.135
T^2	1.597	.386	.144
H^2	.963	.754	.528
R^2	.826	.789	.763
$\hat{\sigma}^2$	.00255	.00308	.00346

4. COMMENTS AND CONCLUSIONS

Biased estimation should be viewed as an important alternative to least squares estimation of the parameters of a multiple linear regression model. Over the last decade biased estimators have been shown to possess valuable theoretical and empirical properties which can be especially advantageous when predictor variables are multicollinear. Although problems persist with the application of biased estimation, their widespread popularity attests to the successful implementation of biased regression methodologies.

Criticisms of biased estimation in regression (e.g. Coniffe and Stone (1973), Draper and Van Nostrand (1979), Smith and Campbell (1980)) serve at least two useful purposes. First they rightfully attack the view that biased estimation provides a panacea for all the problems inherent in regression analyses. In particular, they deplore the simplistic view that one can alleviate the difficulties stemming from multicollinear data sets by, for example, merely selecting a value of the ridge parameter. Second, they focus attention on unresolved theoretical questions associated with biased estimators. Just as Theobald (1974) provided a solution to one such question, further research can conceivably make progress in answering others.

This paper has attempted to illuminate the benefits and deficiencies of biased estimation with special reference to the analysis of regression data. Although criticisms have been levied at the current incomplete theoretical justification for biased estimation, attention has been directed in this paper to the clear advantages of three biased estimators over least squares on the emissions data. In spite of problems with the application of the biased estimators - problems which are acknowledged and illustrated -

all three biased estimators result in coefficient estimates which are more reasonable than least squares. The ultimate criteria for measuring the efficacy of a regression estimator, one can argue, are ones that are impossible to precisely define and can only be subjectively measured: do the coefficient estimates make sense from the physical nature of the problem and does the final prediction equation predict accurately enough? Biased regression estimates often satisfy both of these criteria.

One final comment on the emissions data is in order. It was selected for use in this paper solely because it clearly illustrates the problems associated with multicollinearities and the effects of the three biased estimators. As was mentioned in Section 2, there are many alternatives to least squares with this data. One alternative is to exclude interaction and/or quadratic terms because they are so highly correlated with the linear ones. Hamaker (1962) provides an example (in a variable selection context) of a data set in which an incorrect theoretical model would be fit if such an alternative was adopted. In addition, all the biased estimators used on the emissions data suggest retaining H^2 in spite of its high correlation with other humidity variables. It is also true that the severe multicollinearities disappear if the linear terms are standardized prior to the formation of interaction and quadratic terms. So long as one recognizes that the model parameters are transformed as well, this is another possible alternative. Again, the failure to consider these alternatives is based on the use of this data set to illustrate properties of the various regression estimators.

5. ACKNOWLEDGMENTS

This research is supported in part by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Contract No. F49620-79-C-0106.

REFERENCES

Bock, M.E., Yancey, T.A., and Judge, G.G. (1973). "Statistical Consequences of Preliminary Test Estimators in Regression." J. Amer. Statist. Assoc., 68, 109-116.

Conniffe, D. and Stone, J. (1973). "A Critical View of Ridge Regression." The Statistician, 24, 181-187.

Dempster, A.P., Schatzoff, M., and Wermuth, N. (1977). "A Simulation Study of Alternatives to Ordinary Least Squares." J. Amer. Statist. Assoc., 72, 77-90.

Draper, N.R. and Van Nostrand, R.C. (1979). "Ridge Regression and James-Stein Estimation: Review and Comments." Technometrics, 21, 451-466.

Gibbons, D.I. (1978). "A Simulation Study of Some Ridge Estimators." Research Publication GMR - 2659, General Motors Research Laboratories, Warren, Michigan.

Gunst, R.F. and Mason, R.L. (1977a). "Biased Estimation in Regression: An Evaluation Using Mean Squared Error." J. Amer. Statist. Assoc., 72, 616-628.

Gunst, R.F. and Mason, R.L. (1977b). "Advantages of Examining Multicollinearities in Regression Analysis." Biometrics, 33, 249-260.

Gunst, R.F., Webster, J.T., and Mason, R.L. (1976). "A Comparison of Least Squares and Latent Root Regression Estimators." Technometrics, 18, 75-83.

Hamaker, H.C. (1962). "On Multiple Regression Analysis." Statistica Neerlandica, 16, 31-56.

Hare, C.T. and Bradow, R.L. (1977). "Light Duty Diesel Emission Correction Factors for Ambient Conditions." Paper No. 770717, Society of Automotive Engineers Off-Highway Vehicle Meeting, MECCA, Milwaukee, Sept. 12-15, 1977.

Hocking, R.R. (1976). "The Analysis and Selection of Variables in Linear Regression." Biometrics, 32, 1-50.

Hoerl, A.E. (1962). "Application of Ridge Analysis to Regression Problems." Chemical Engineering Progress, 58, 54-59.

Hoerl, A.E. and Kennard, R.W. (1970a). "Ridge Regression: Biased Estimation for Nonorthogonal Problems." Technometrics, 12, 55-67.

Hoerl, A.E. and Kennard, R.W. (1970b). "Ridge Regression: Application to Nonorthogonal Problems." Technometrics, 12, 69-82.

Hoerl, A.E., Kennard, R.W., and Baldwin, K.F. (1975). "Ridge Regression: Some Simulations." Comm. in Statist., 4, 105-123.

Kendall, M. (1957). A Course in Multivariate Analysis. London:Griffin.

Marquardt, D.W. (1970). "Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation." Technometrics, 12, 591-612.

Marquardt, D.W. and Snee, R.D. (1975). "Ridge Regression in Practice." American Statistician, 29, 3-20.

Massy, W.F. (1965). "Principal Component Regression in Exploratory Statistical Research." J. Amer. Statist. Assoc., 60, 234-256.

McDonald, G.C. and Galarneau, D.I. (1975). "A Monte Carlo Evaluation of Some Ridge-Type Estimators." J. Amer. Statist. Assoc., 70, 407-416.

Silvey, S.P. (1969). "Multicollinearity and Imprecise Estimation." J. Roy Statist. Soc., B, 31, 539-552.

Smith, G. and Campbell, F. (1980). "A Critique of Some Ridge Regression Methods." J. Amer. Statist. Assoc., 75, (to appear).

Theobald, C.M. (1974). "Generalizations of Mean Square Error Applied to Ridge Regression." J. Roy. Statist. Soc., B, 36, 103-106.

Webster, J.T., Gunst, R.F., and Mason, R.L. (1974). "Latent Root Regression Analysis." Technometrics, 16, 513-522.

White, J.W. and Gunst, R.F. (1979). "Latent Root Regression: Large Sample Analysis." Technometrics, 21, 481-488.

Wichern, D.W. and Churchill, G.A. (1978). "A Comparison of Ridge Estimators." Technometrics, 20, 301-312.